Distributed and Peer-to-Peer Data Mining for Scalable Analysis of Data from Virtual Observatories

Hillol Kargupta, University of Maryland, Baltimore County
Kirk Borne, George Mason University
Chris Giannella, Loyolla College

Road Map

- Motivation
- Distributed Data Mining
- Current Research
- Preliminary Results
- Future Directions

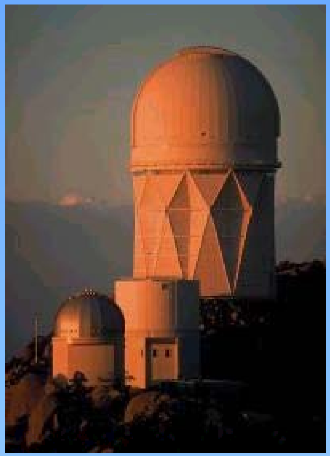
Data Mining and Distributed Data Mining

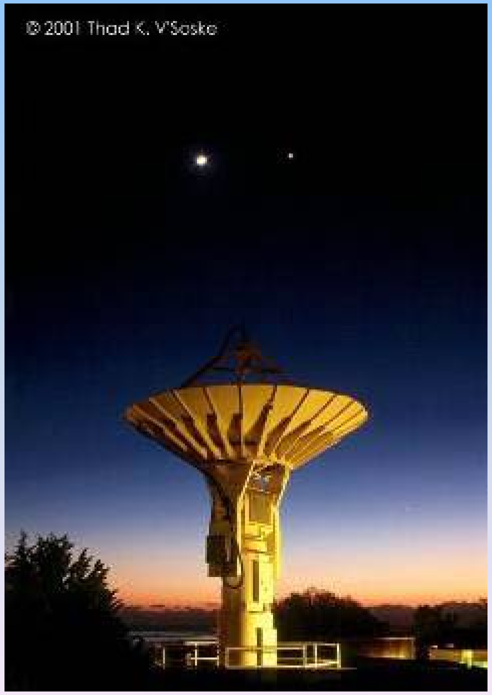
- Data Mining: Scalable analysis of data by paying careful attention to the resources:
  - computing,
  - communication,
  - storage, and
  - human-computer interaction.
- Distributed data mining (DDM): Mining data using distributed resources.

Future of Astronomy Data Processing Environment

- Multiple data sources
- Heterogeneous distributed computing environment
- Increasing number of users; scientific communities forming peer-to-peer networks
- Increasing demand for faster response time
- High throughput data streams

Multiple Data Sources





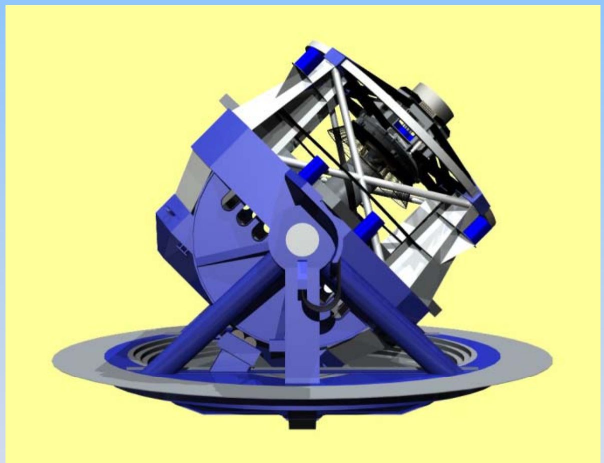
Why Multiple Sources?

And because ...

- Astrophysical correlations often relate object parameters that are measured in multiple wavebands, hence different surveys:
  - Optical images, photometry, and galaxy redshifts from SDSS (Sloan Digital Sky Survey)
  - Near-infrared (NIR) images and photometry from the 2-Micron All-Sky Survey (2MASS)

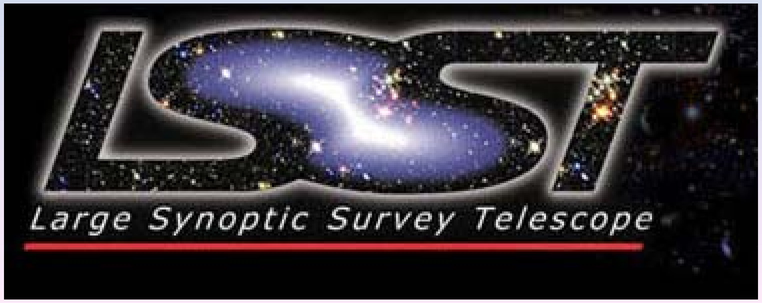
Astronomy Data Streams

Future sky-surveys are expected to produce large quantities of new observational data nightly.



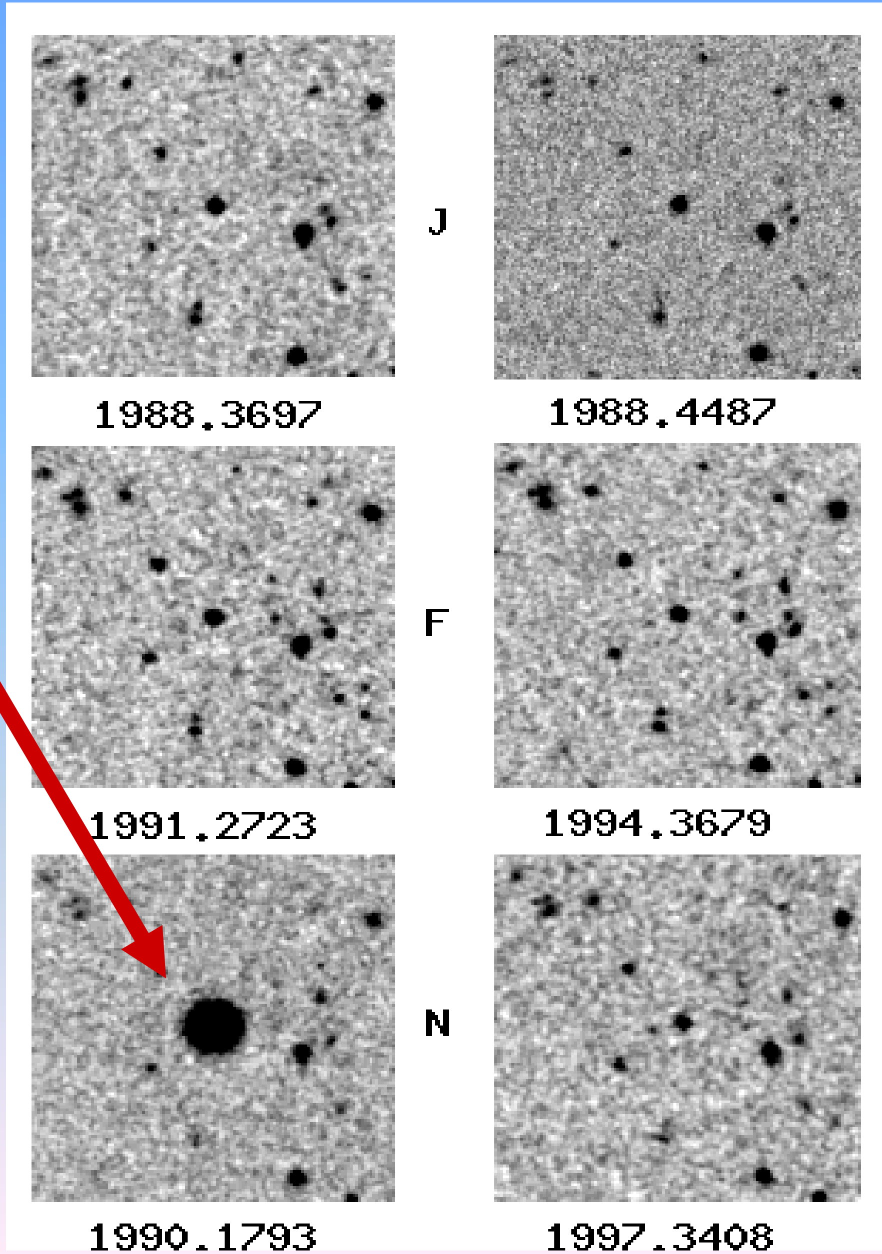
e.g., LSST is expected to produce

- ≈ 30 TB of data per night for 10 years
- ≈ 20 PB final astronomical object catalog



Alert Processing

- Potentially 10^5 alerts per night. An alert is any observed change (e.g., object brightness, object location, new object appearance, etc.) in the night sky.
- Rapid, appropriate follow-up observations are essential for short-lived phenomenon to determine what it is, what caused the sudden variation in its properties, and how does it relate to other objects.



Alert Classification

- Classify alerts (events) to facilitate appropriate, rapid response.
  - time-criticality, trigger appropriate follow-up observation resources, etc..
- Produce tagged alerts for the astronomy community
  - tags will include classifications
  - classifications will require distributed data mining across the NVO (i.e., feature set drawn from multiple astrophysical data repositories)
  - viewable in near real-time over the WWW

Distributed and Peer-to-Peer Computing Environment

Distributed computing environments (community of users)

High performance grid computing

Overall Project Goal

- Develop scalable data mining techniques for quickly sifting through distributed data streams

- Then download data if necessary for more in-depth analysis based on the feedback from Step 1.

Project Methodology

- Identify a set of important astronomy data mining tasks
- Develop distributed and P2P stream mining algorithms for these tasks
  - Principal component analysis
  - Outlier detection
  - Classification

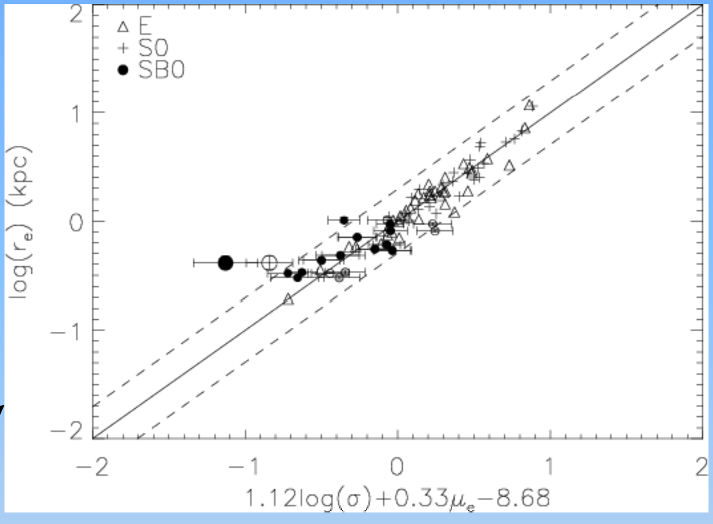
Mining Multiple Surveys Across the NVO – Steps 1 and 2

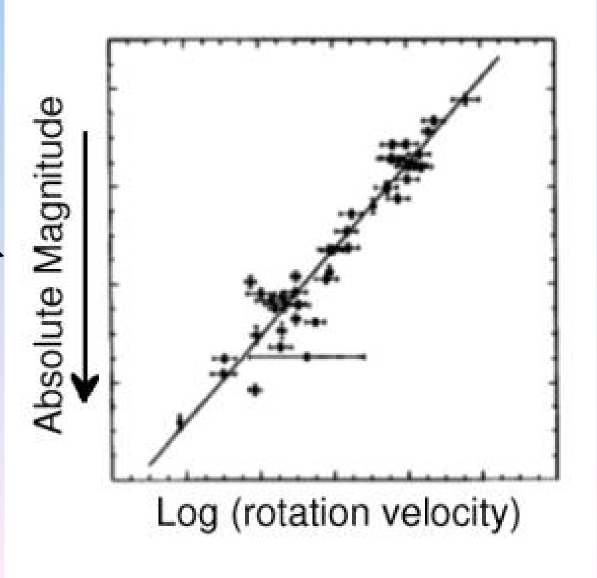
- 1. Identify the research domain and the research problem: *explore large astronomical data sets for new classes of objects through discovery and analysis of correlations*

- 2. Download, cross-match, and preprocess data from multiple distributed astronomical repositories (NVO)
  - e.g., SDSS, 2MASS

Mining Multiple Surveys Across the NVO – Step 3

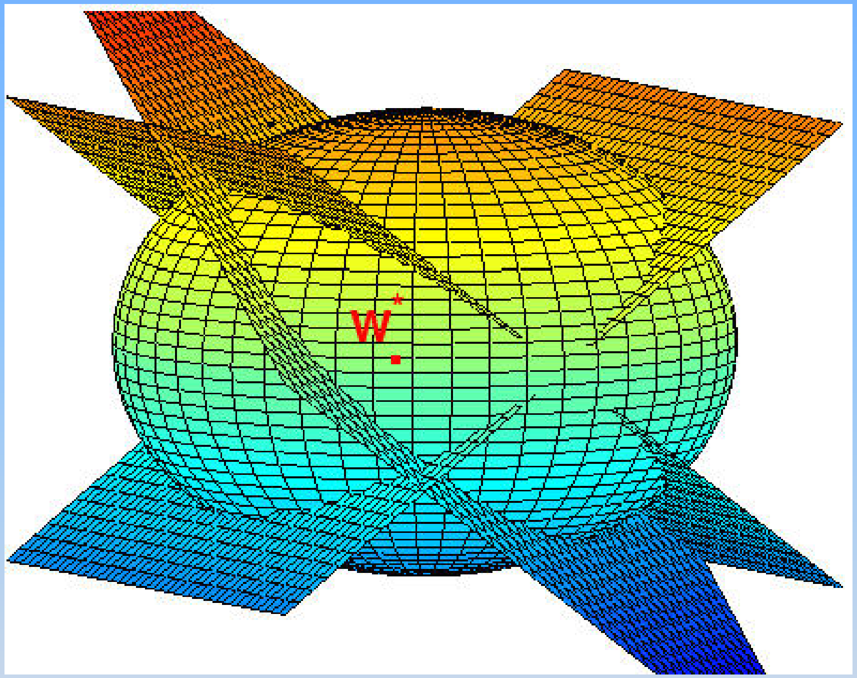
- 3. Apply data mining to address questions more readily than could be done using individual catalogs
  - e.g., Are there systematic variations in known astrophysical correlations (e.g., the fundamental plane of elliptical galaxy properties, or the Tully-Fisher relation of spiral galaxies) as a function of local galaxy density?





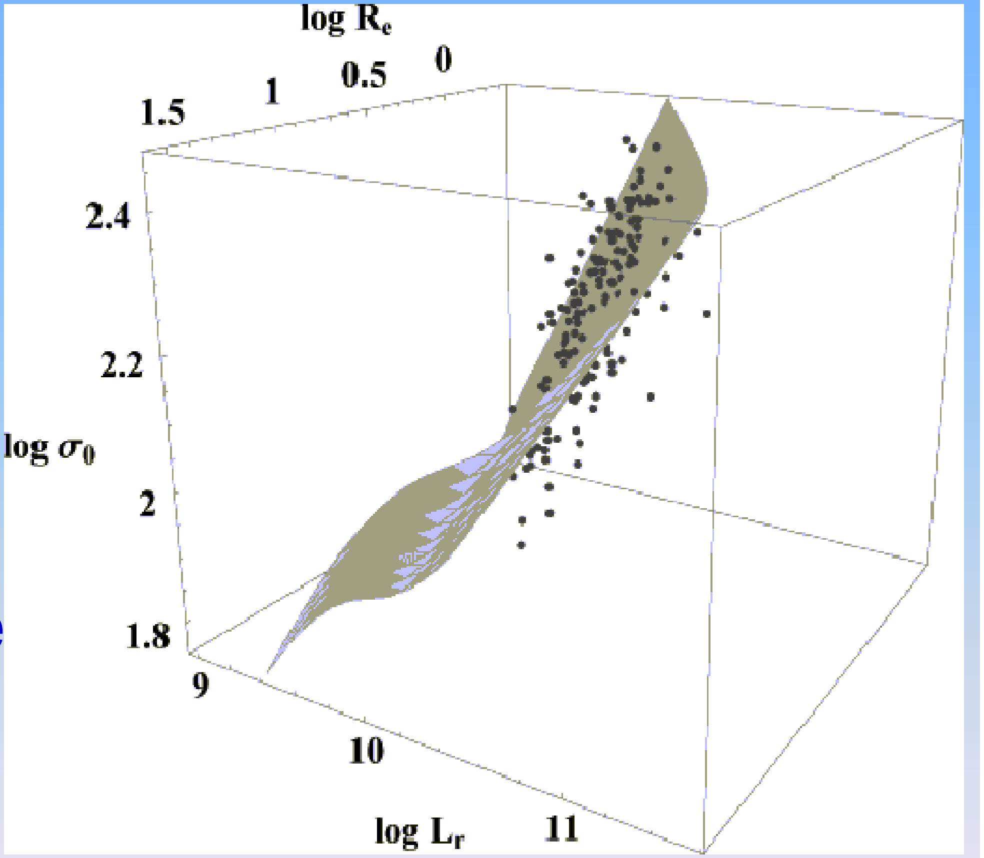
Mining Multiple Surveys Across the NVO – Step 3

- 3. i.e., Does the normal vector of the principal plane (or the variance captured in the principal components) of complex multi-parameter relationships vary systematically with local galaxy density?



The Fundamental Plane of Elliptical Galaxies

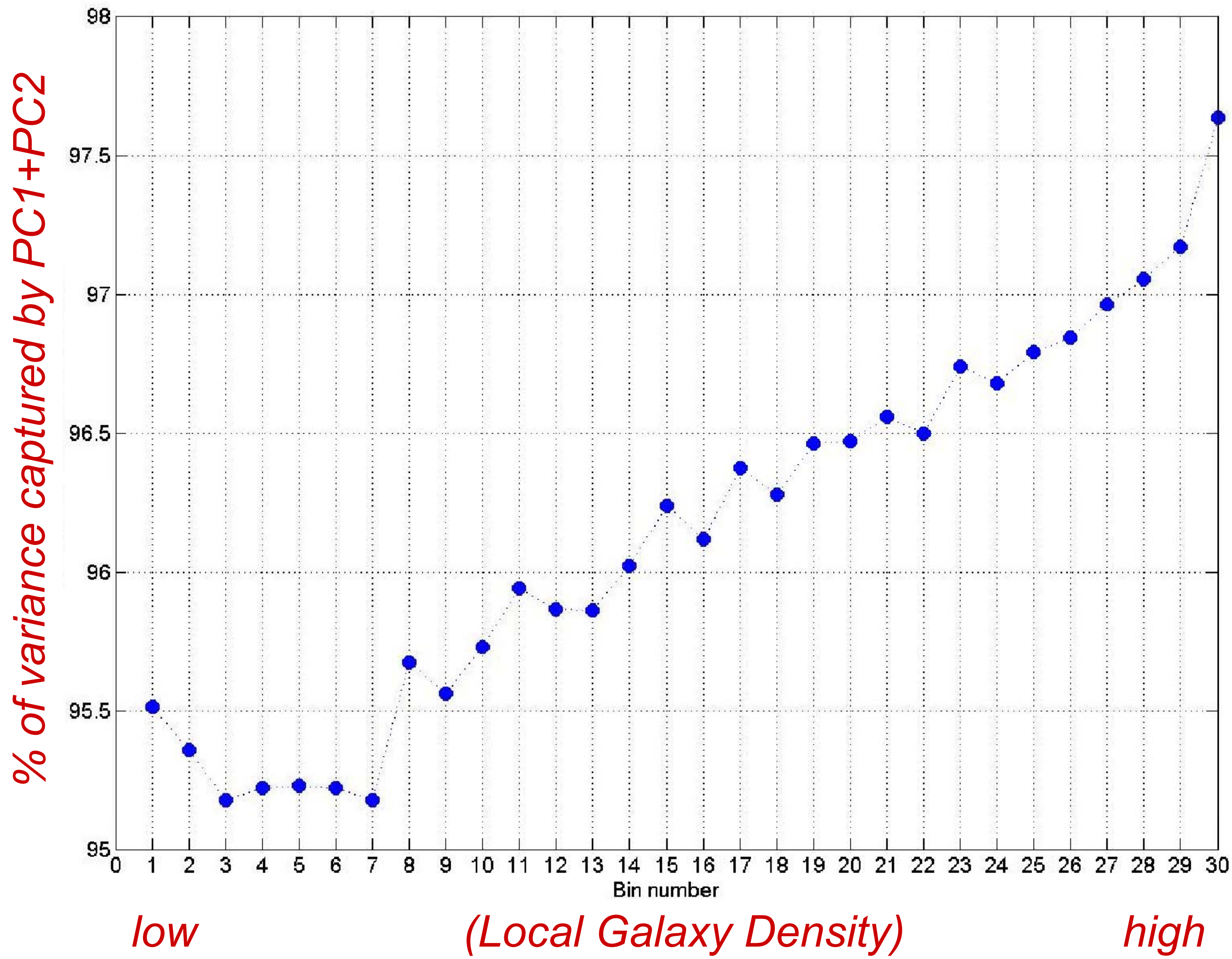
- The **fundamental plane** for elliptical galaxies tracks the correlation between the effective radius, average surface brightness, and central velocity dispersion.
- With this correlation, one can determine the distance to galaxies, which is a critical but difficult task in astronomy.



Preliminary Experiment

- Produced a 156,000 cross-matched galaxy dataset with attributes from SDSS and 2MASS
  - SDSS: velocity dispersion, Pertrosian I band angular effective radius, redshift
  - 2MASS: K band mean surface brightness
- Partitioned into bins w.r.t. local galaxy density *ρ*
  - Estimated *ρ* using Delaunay tessellation methods
  - The local density around a selected galaxy is inversely proportional to the local volume that contains only that one galaxy – measured by the Delaunay tessellation
- Estimated the fundamental plane parameters for each bin (e.g., variance captured in the first two principal components, as *f(ρ)*)

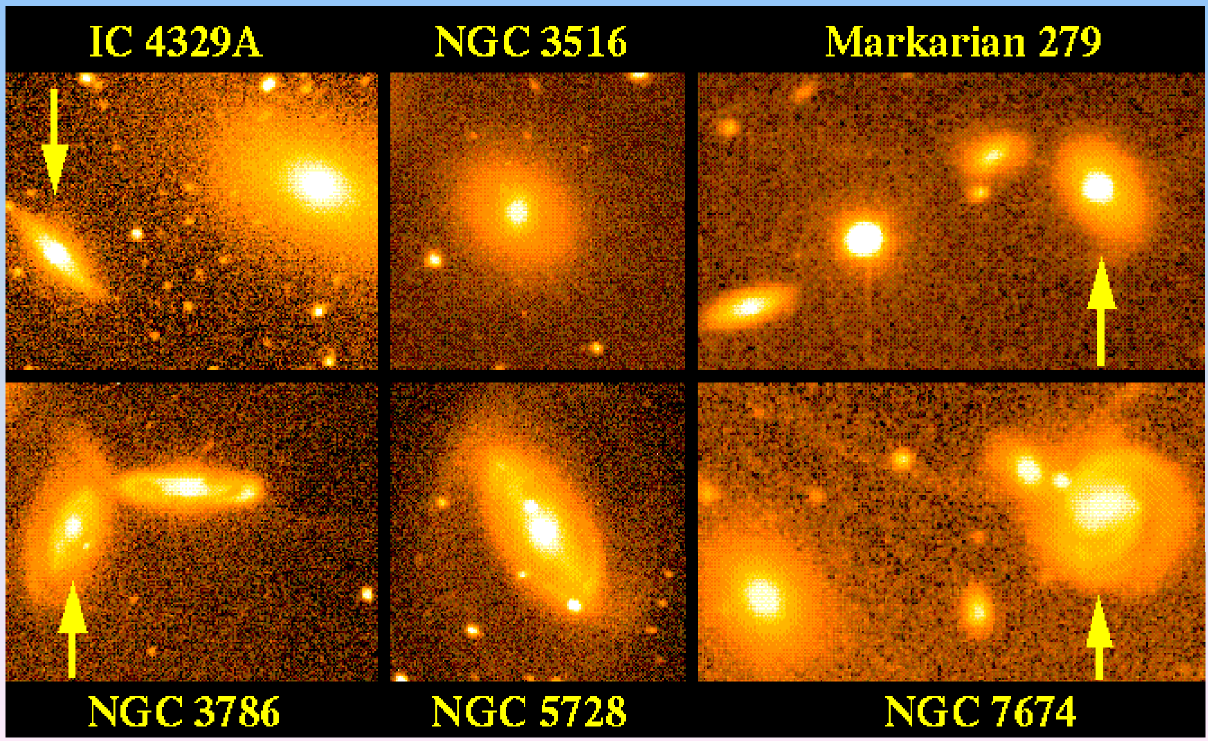
Results



% of variance captured by PC1+PC2
98
97.5
97
96.5
96
95.5
95
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Bin number
low
(Local Galaxy Density)
high

Potential Significance

- This systematic trend shows that the strength of the correlation (fundamental plane) grows as the local galaxy density increases.
- This indicates that the dynamical properties (hence formation scenarios) of elliptical galaxies become increasingly uniform in the densest environments.

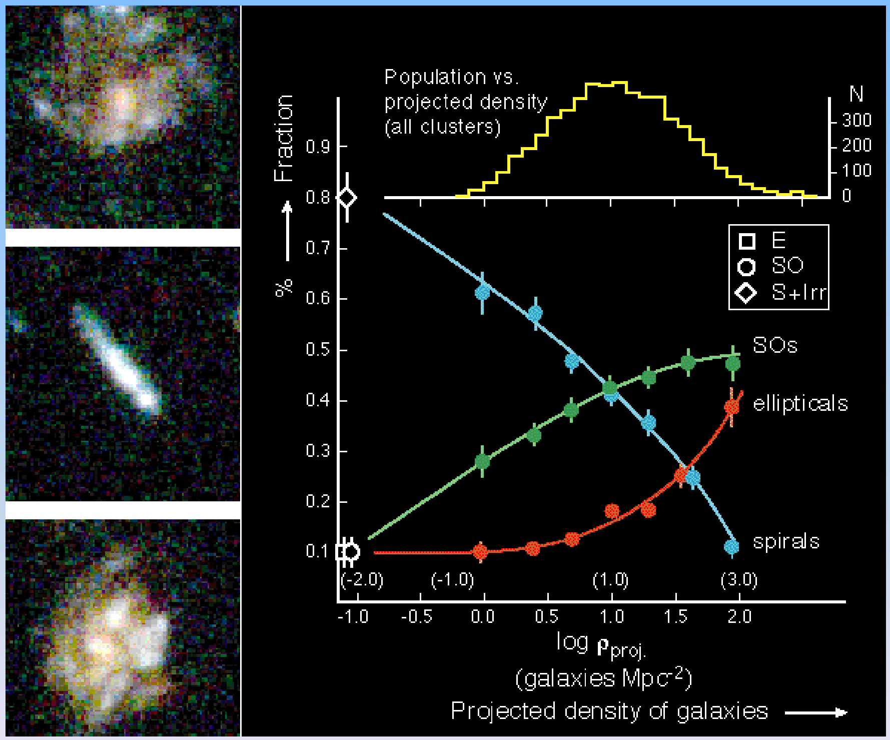


Distributed and P2P Algorithm Design

- Distributed and P2P PCA
- Can be reduced to distributed inner product computation
- Asynchronous local algorithms for distributed inner product computation
  - Exact
  - Approximate

Fundamental Plane Problem: Further Refinement

- Further homogenization of the cross-matched galaxy sample
  - e.g., to filter out bulge-dominated spirals = non-ellipticals
  - will help to isolate the trends found here
  - i.e., to determine if the trends that we found are a new discovery, or are they a manifestation of the well known morphology-density relationship of galaxies?



Future Work

- Expand the scope of the scientific research; identify other scientific tasks.
- Develop distributed algorithms

Recent Awards and Sample Publications

- K. Bhaduri and H. Kargupta. (2008). An efficient local Algorithm for Distributed Multivariate Regression in Peer-to-Peer Networks. 2008 SIAM Data Mining Conference Paper selected for “Best of SDM’08” issue (one of the top six papers in SDM’08)

- H. Dutta, C. Giannella, K. Borne and H. Kargupta. (2007). Distributed Top-K Outlier Detection from Astronomy Catalogs using the DEMAC System. Proceedings of the SIAM International Conference on Data Mining, Minneapolis, USA, April 2007.